Bringing visibility to food security data results
-- *Harvests of PRAGMA and RDA*

Quan (Gabriel) Zhou [1], Venice Juanillas [2], Ramil Mauleon [2],
Jason Haga [3], Beth Plale [1]
Indiana University, USA [1] / IRRI, Philippines [2] / AIST, Japan [3]

Introduction

Globally unique Persistent IDs (PIDs) promise to make research data easier to manage. The RDA Recommendations, Persistent Identifier Types (PIT) and Data Type Registry (DTR), are enabling solutions. A group of us in the Pacific Rim Applications and Grid Middleware Assembly (PRAGMA), a strong network of member institutions around the Pacific Rim, adopted PIT and DTR into a application that runs Galaxy workflows on PRAGMA VMs running on the PRAGMA Cloud Testbed.

The user group benefiting from our work are members of the International Rice Research Institute (IRRI) in Manila, Philippines. The services run at the National Institute for Advanced Industrial Science and Technology (AIST), Japan. The development work was carried out by Indiana University Data To Insight Center and AIST.

IRRI Tassel5 Workflow

Application is Tassel5 pipeline [1] provided through Galaxy interactive workflow system [2].

Application enables GWAS analysis of researchers' own phenotyping data with robust SNP data subset of 3K Rice Genomes (3KRG) [3] using common analysis framework

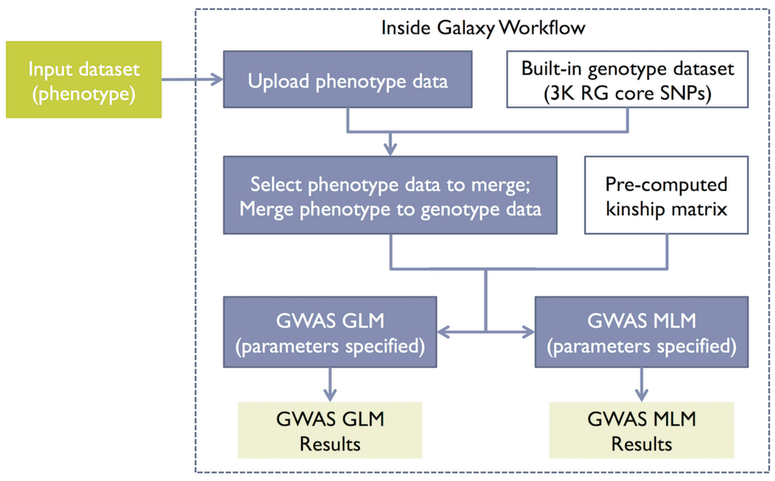

Fig. IRRI Rice Genomics Tassel5 Galaxy Workflow

Persistent identification services interact with workflow system to automatically harvest digital objects (DO's).

Offers step to reproducibility, user transparency, and minimum instrumentation to Tassel5 pipeline.

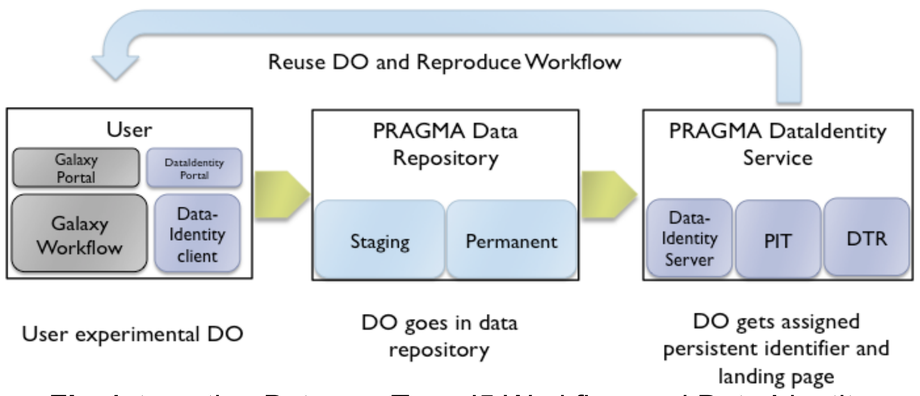

Fig. Interaction Between Tassel5 Workflow and Data Identity Service

Framework

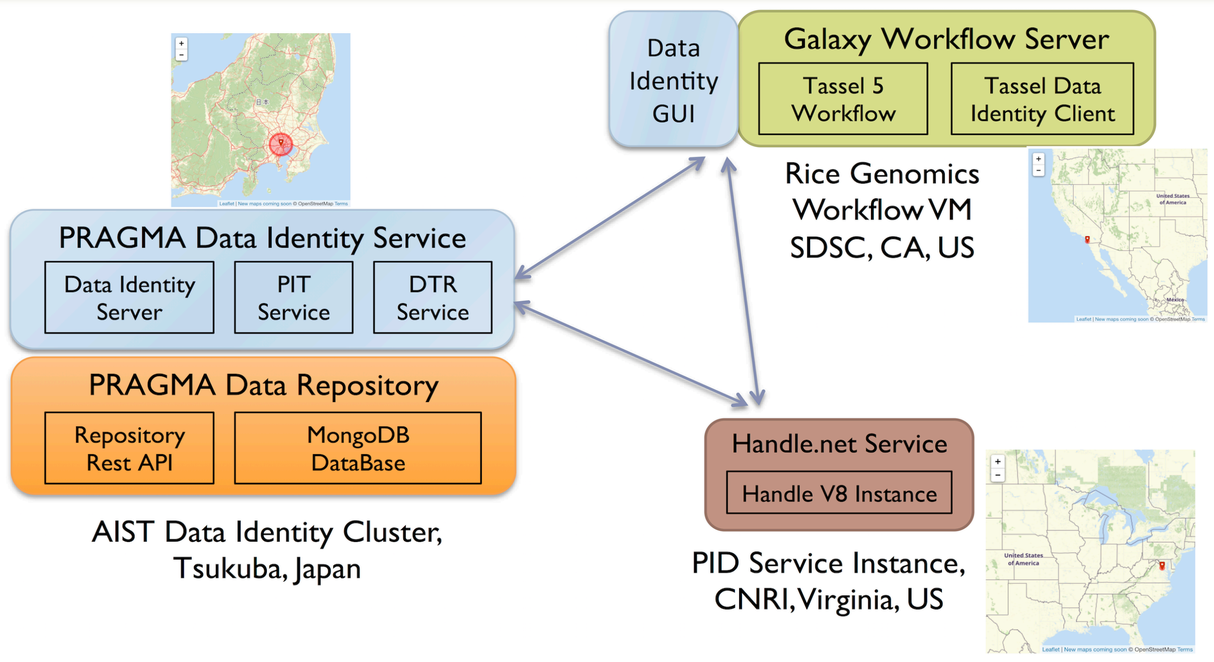

Fig. Service Deployment Diagram in AIST, SDSC and CNRI

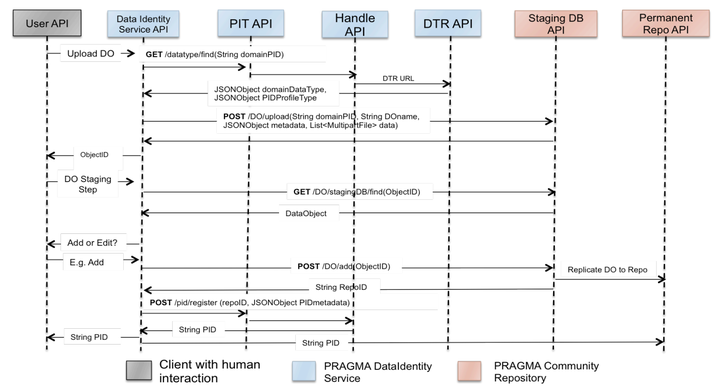

Fig. DO Upload Timeline Diagram

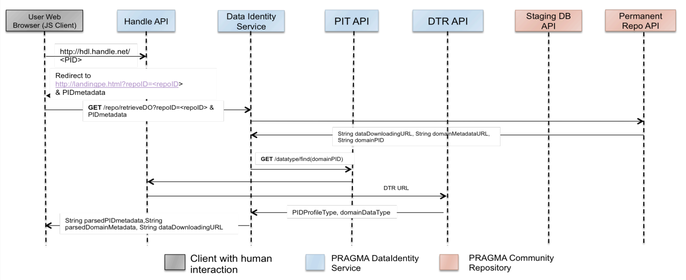

Fig. DO Retrieval Timeline Diagram

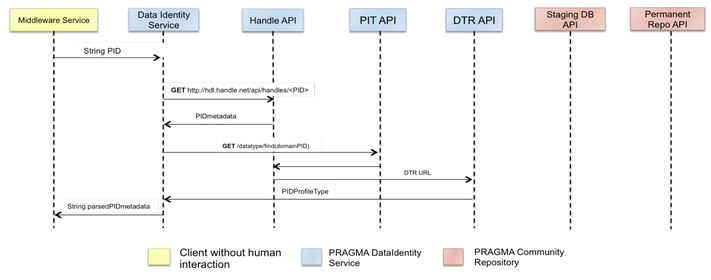

Fig. Middleware Service Timeline Diagram

Success to Date

The PRAGMA Data Identity Services is a user transparent means of harvesting DOs from applications and assignment of PIDs to scientific outcomes.

Successes to date include:

- Has been reviewed by core members of the rice genomics team
- User group study of members of IRRI rice genomics community planned to occur Fall 2016
- Software is stable. Early version of the software available on Github.
- Built with default PID information types and metadata

Acknowledgement
This project funded in part by Research Data Alliance/US through grant from MacArthur Foundation; by NSF grant # OCI-#1234983, and by funding from AIST ICT International Team.

Future Work

Data identity services and PRAGMA repository can be used in other applications running on and off the PRAGMA Cloud testbed.

Next steps:

- User interface and hardening over Fall 2016 for more robust operation.
- Refine metadata types based on user group study feedback
- Extend data server (mongoDB) with basic preservation capabilities
- Install services at US National Data Service (NDS)
- Serve as basis for US testbed
- Evaluate provenance capture

References

[1] Bradbury PJ, et al. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.
[2] Mauleon, R. et al., 2012. IRRI GALAXY: bioinformatics for rice. ISCB-Asia/SCCG, 2012, Shenzhen, China
[3] Li J.Y., Wang J., and Zeigler R.S., 2014. The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience, 2014. 3: 8. DOI: 10.1186/2047-217X-3-8
[4] Weigel, T. et al., 2013. A Framework for Extended Persistent Identification of Scientific Assets. Data Science Journal, 12, pp.10–22. DOI: http://doi.org/10.2481/dsj.12-036